How Diachronic Text Corpora Affect Context based Retrieval of OOV Proper Names for Audio News

Imran Sheikh, Irina Illina, Dominique Fohr

How Diachronic Text Corpora Affect Context based Retrieval of OOV Proper Names for Audio News

Imran Sheikh, Irina Illina, Dominique Fohr
Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
{imran.sheikh, irina.illina, dominique.fohr}@loria.fr

Abstract

Out-Of-Vocabulary (OOV) words missed by Large Vocabulary Continuous Speech Recognition (LVCSR) systems can be recovered with the help of the topic and lexical context of the OOV words captured from a diachronic text corpus. In this paper we investigate, how the choice of documents in the diachronic text corpora affects the retrieval of OOV Proper Names (PNs) relevant to an audio document. We first present our French diachronic broadcast news datasets which highlight the motivation of our study on OOV PNs. Then the effect of using diachronic text data from different sources and different time span is analysed. With OOV PN retrieval experiments on French broadcast news videos, we conclude that diachronic corpus with text from different sources leads to better retrieval performance than relying on text from single source or from a longer time span.

Keywords: oov, proper names, lvcsr, diachronic corpus

1. Introduction

Diachronic news is characterised by different topics which change with time, leading to a change in the linguistic content and vocabulary. As a result, a typical problem faced by *Large Vocabulary Continuous Speech Recognition* (LVCSR) systems processing such diachronic audio news is *Out-Of-Vocabulary* (OOV) words. In previous works it has been observed that majority of OOV words are PNs. PN percentage in OOV words being reported as :56% in (Qin, 2013), 66% in (Parada et al., 2011), 57.6% in (Palmer and Ostendorf, 2005), 70% in (Allauzen and Gauvain, 2005), 72% in (Béchet et al., 2000). On the other hand, *Proper Names* (PNs) in audio news are of prime importance for content based indexing and browsing applications. Our work is closely related to recognition and recovery of OOV PNs i.e., PNs which appear in diachronic audio but are not present in the LVCSR vocabulary and cannot be recognised by the LVCSR system.

To recognise OOV PNs in test audio documents, we rely on new PNs extracted from collections of diachronic text news from the internet (referred as *diachronic corpus*). Instead of using all these new PNs to recognise/recover the *target OOV PNs*[1], it would be efficient to use only the relevant new PNs. In (Sheikh et al., 2015a), we proposed methods to reduce the long list of OOV PNs from the diachronic corpus to a smaller list of OOV PNs which are relevant to the test audio document. We refer to this task as *retrieval of OOV PNs relevant to an audio document*. To achieve this, we used the topic and lexical context of the audio document and OOV PNs.

The topic and lexical context, which enable the retrieval of relevant OOV PNs, are derived from (and are dependent on) the diachronic corpus. In this paper we try to investigate some characteristics of the diachronic corpus which can affect the performance of retrieval of OOV PNs. In our previous works (Sheikh et al., 2015a; Sheikh et al., 2015b) we have presented the problem of less frequent OOV PNs in the diachronic corpus, discussed topic model variations and proposed methods to handle the less frequent OOV PNs. In contrast, our study in this paper is around selection of documents for the diachronic corpus, which is used to model context. For instance we try to analyse the effect on the retrieval performance when: (a) diachronic corpus is built with text from different originating sources (b) diachronic corpus from a single source is supplemented with text resources only for the less frequent OOV PNs (c) diachronic corpus timeline extends far beyond test dataset timeline. Following our earlier works we will focus on retrieval of OOV PNs in French broadcast news videos.

The rest of the paper is organised as follows. In Section 2. we briefly discuss our approach to retrieve OOV PNs relevant to an audio document using probabilistic topic models. In Section 3. we present French diachronic broadcast news datasets used in our experiments and in Section 4. we present the different configurations of diachronic corpora used in our study. Section 5. presents the experiments and results, followed by discussion in Section 6..

2. OOV PN Retrieval using Topic Models

As mentioned earlier, our task is to retrieve OOV PNs relevant to an audio document. To achieve this, we rely on new PNs extracted from collections of diachronic text news from the internet which we refer as diachronic corpus. Topic models are trained using a diachronic text corpus as training corpus to learn relations between words, latent topics and OOV PNs. During test, the audio news document is transcribed by the LVCSR with the standard vocabulary. *In-Vocabulary* IV words (including IV PNs) hypothesised by the LVCSR are then input to the topic model to infer

[1] Ideally new PNs extracted from collections of diachronic text news are OOV PNs with respect to the LVCSR vocabulary. However all new PNs are not present in the test set audio documents. Hence we use the term *target OOV PNs* to refer to the OOV PNs actually present in the test set audio documents. The general term 'OOV PNs' is used to refer to new PNs.

Table 1: Broadcast news diachronic datasets

	L'Express (LX)	*Le Figaro* (FIG)	*L'Express* + *Le Figaro* (LX+FIG)	*L'Express* (LX-18m)	*Euronews* (Dev)	*Euronews* (Test)
Type of Documents	Text	Text	Text	Text	Text	Video
Time Period	Jan - Jun 2014	Jan - Jun 2014	Jan - Jun 2014	Jul 2013 - Dec 2014	Jan - Jun 2014	Jan - Jun 2014
Number of Documents[1]	45K	59K	104K	142K	3.1K	3K
Vocabulary Size (unigrams)[2]	150K	140K	180K	270K	42K	45K
Corpus Size (approx. word count)	24M	18M	42M	70M	550K	700K
Number of PN unigrams[2]	57K	51K	80K	104K	12K	11K
Total PN count	1.45M	1.3M	2.7M	4.2M	54K	42K
Number of OOV unigrams[3]	12.4K	11.9K	24.4K	37.1K	4.9K	4.3K
Documents with OOV[3]	32.3K	36.4K	73K	109K	2.25K	2.2K
Total OOV count[3]	141K	142K	320K	509K	9.1K	8K
Number of OOV PN unigrams[3]	9.3K	8.8K	18.4K	28.2K	3.4K	3.1K
Documents with OOV PN[3]	26.5K	30K	61.3K	93.5K	1.9K	1.9K
Total OOV PN count[3]	107K	103K	243K	388K	6.9K	6.2K

[1]K denotes *Thousand* and M denotes *Million*, [2] unigrams occurring less than two times are ignored
[3]unigrams occurring in less than three documents ignored, documents with more than 20 and less than 500 terms
Note: (a) OOV and OOV PN statistics are post term-document filtering (b) [2] and [3] does not apply for Euronews

the topic distribution of the test audio document. With the test document topic distribution we retrieve a list of most relevant OOV PNs for the test document.

Following our approach in (Sheikh et al., 2015a; Sheikh et al., 2015b), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is used to model topics of the diachronic corpus. To model topics in the diachronic corpus of text documents, we choose a fixed vocabulary, the number of topics T and symmetric Dirichlet priors. Topic model parameters θ and ϕ are estimated using Gibbs sampling algorithm (Griffiths and Steyvers, 2004). Topics learned by LDA are used to retrieve OOV PNs relevant to the test document. Let us denote LVCSR hypothesis of test document by h and OOV PNs in diachronic corpus by $\tilde{v}_x$. In order to retrieve OOV PNs, we calculate $p(\tilde{v}_x|h)$, for each $\tilde{v}_x$ and then treat it as a score to rank OOV PNs relevant to h. With the words observed in h, the latent topic mixture $[p(t|h)]_T$ can be inferred by re-sampling from the word-topic distribution ϕ learned during training (Griffiths and Steyvers, 2004). Given $p(\tilde{v}_x|t) = \phi_{vt}$, the likelihood of an OOV PN ($\tilde{v}_x$) can be calculated as: $p(\tilde{v}_x|h) = \sum_{t=1}^{T} p(\tilde{v}_x|t)\, p(t|h)$.

3. Diachronic Broadcast News Datasets

Table 1 presents realistic broadcast news diachronic datasets which will be used as the training and test sets in our study. These datasets also highlight the motivation for handling OOV PNs. The datasets are collected from three sources: (a) French newspaper *L'Express* (http://www.lexpress.fr/) (b) French newspaper *Le Figaro* (http://www.lefigaro.fr/) (c) French website of the *Euronews* (http://fr.euronews.com/) television channel. The *L'Express* and *Le Figaro* datasets contain text news whereas the Euronews dataset contains text news as well news videos and their text transcriptions. In our study the *L'Express* and *Le Figaro* datasets will be used as diachronic corpora to train topic models, in order to infer the OOV PNs relevant to Euronews videos.

TreeTagger[2] (Schmid, 1994) is used to automatically tag PNs in the text. The words and PNs which occur in the lexicon of our *Automatic News Transcription System* (ANTS) (Illina et al., 2004) are tagged as IV and remaining PNs are tagged as OOV. ANTS lexicon is based on news articles until 2008 from French newspaper *LeMonde* and contains about 123K unique words. As shown in Table 1, 72% (3.1K out of 4.3K) of OOV words in the *Euronews* video dataset are PNs and about 63% (1.9K out of 3K) of the videos contain OOV PNs. An important statistic *target OOV PN coverage* is not shown in Table 1. We use the term target OOV PN coverage to refer to the percentage of the target OOV PNs in Euronews videos which can be recovered with a diachronic corpus. The target OOV PN coverage for each of the diachronic text corpus is found to be: for LX 42%, for FIG 40%, for LX+FIG combined 52% and for LX-18m 54%.

4. Diachronic Corpus Configurations used in our Study

The topic context, which enables the retrieval of relevant OOV PNs, are dependent on the diachronic corpus used to model topic context. We try to study the effect of selection of documents for the training/diachronic corpus. In particular we study the following configurations of the diachronic corpus.

(A) Documents containing OOV PNs[3] and from the same time period as the test set, e.g. *L'Express* documents

[2]http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[3]From our experiments we observed that adding non OOV PN documents to the diachronic corpus does not give significant improvement in the OOV PN retrieval performance.

containing OOV PNs and corresponding to 6 months of the *Euronews* video test set.

(B) Documents coming from two different originating sources, e.g. *L'Express* and *Le Figaro*.

(C) Documents coming from a time period much longer than the test set (e.g. *L'Express* documents from 18 months for the *Euronews* video test set).

(D) Documents with OOV PNs collected from one source and then for the less frequent OOV PNs new documents are additionally collected from another source. As discussed in our previous works (Sheikh et al., 2015a; Sheikh et al., 2015b), retrieval of less frequent OOV PNs has a poor performance because there is not enough data to learn their topic distribution. This problem motivates us to try this configuration.

5. Experiment Setup

5.1. Corpora Setup

In our experiments, we use the datasets presented in Section 3.. The *L'Express* and *Le Figaro* datasets will be used as diachronic corpus to train the topic models. Audio news extracted from the Euronews video dataset is used as the test set. The different configurations for diachronic corpus discussed in Section 4. will be studied with the *L'Express* and *Le Figaro* datasets. Throughout the experiments and discussions, LX, FIG and LX+FIG denote diachronic corpora with documents from *L'Express*, *Le Figaro* (both from Jan 2014 - Jun 2014) and a combination of the two respectively. These correspond to configurations A and B in Section 4. LX-18m denotes diachronic corpus with documents of *L'Express* from Jul 2013 - Dec 2014 and corresponds to configuration C in Section 4. LX+rFIG denotes documents of LX (*L'Express*, Jan 2014 - Jun 2014) supplemented with those documents of FIG (*Le Figaro*, Jan 2014 - Jun 2014) which contain OOV PNs which occur less than 10 times in LX. This corresponds to configuration D in Section 4. Target OOV PN coverage for LX+rFIG is 49%.

5.2. LVCSR processing

The ANTS (Illina et al., 2004) LVCSR system is used to perform automatic segmentation and speech-to-text transcription of the test audio news. The automatic transcriptions of the test audio news obtained by ANTS have an average *Word Error Rate* (WER) of 40% as compared to the reference transcriptions[4] available from *Euronews*.

5.3. Topic Models

For training topic models, diachronic corpus words are lemmatised and filtered by removing PNs and non PN words occurring less than 3 times. Additionally a stoplist of common french words and non content words which do not carry any topic-related information is used. Moreover, a POS based filter is employed to choose words tagged as PN, noun, adjective, verb and acronym. PNs not present in the ANTS LVCSR lexicon are tagged as OOV PNs. LDA topic model is trained with this filtered vocabulary. We trained 100, 200, 300, 400 and 500 topics for each of the corpora configurations and first evaluated the performance on our *Euronews* (text) development set. Figure 1, shows the effect of number of topics on the OOV PN retrieval performance. The comparison in Figure 1 is in terms of *Recall@100* and *Map@100* (Manning et al., 2008), which represent the Recall and MAP values obtained by considering the top 100 retrieved OOV PNs. Higher number of topics are more favourable for larger corpus and we can observe the same from Figure 1. LX-18m and LX+FIG continue to give better performance with increased number of topics but for LX and FIG the recall does not improve beyond 300 topics. The MAP however still improves for LX and Fig and we will discuss more about this difference in Section 6.1. As each diachronic corpus configuration will perform its best for a certain number of topics, choosing a fixed number of topics for analysis of corpora of different sizes may not be appropriate. However, since our study is focused on effects of corpus selection on the (OOV PN) retrieval performance, we will continue our analysis on the test set with each diachronic corpus modelled with 300 topics. While we will refer back to Figure 1 to support and improve relevant observations whenever required.

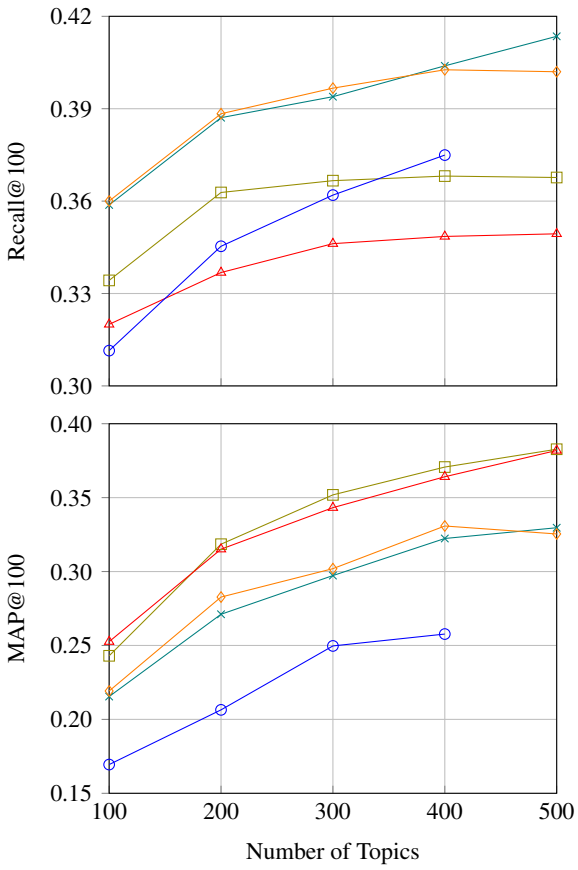

Figure 1: Recall@100 and MAP@100 for OOV PN retrieval on *Euronews* text (Dev set) with different number of LDA topics trained on the diachronic corpora. (LX, FIG, LX+FIG, LX+rFIG, LX-18m)

[4]these are approximate transcriptions; on a set of 10 manually transcribed audio files we obtained a WER of 33.8%

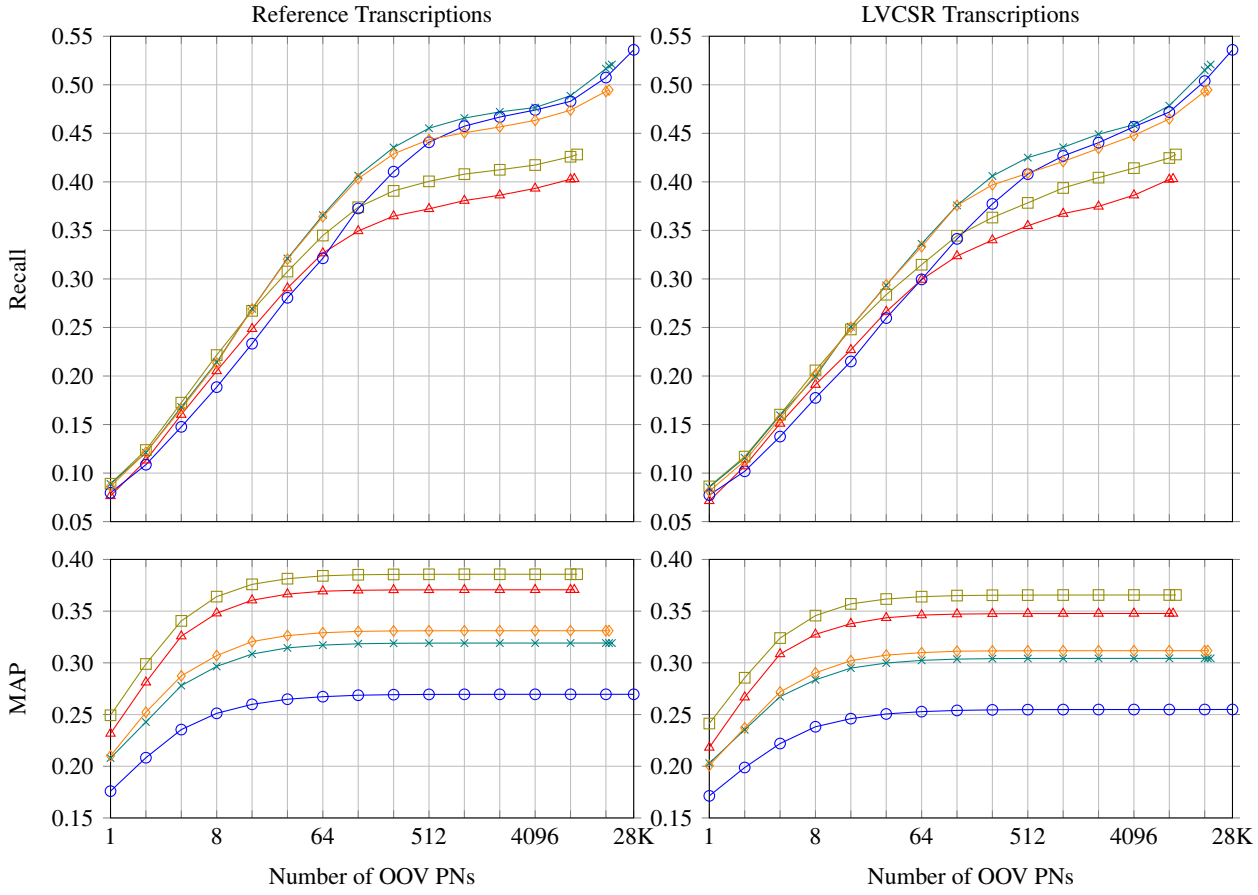

Figure 2: Recall and MAP for OOV PN retrieval on *Euronews* news videos with different diachronic corpora for training LDA topic model. (LX, FIG, LX+FIG, LX+rFIG, LX-18m)

6. Discussion

Figure 2 helps us to analyse and compare the effects of diachronic corpus configurations A, B, C and D discussed in Section 4. Figure 2 shows the *Recall* and *Mean Average Precision* (MAP) (Manning et al., 2008) performance of retrieval of OOV PNs. The performance on both reference (left) and LVCSR (right) transcriptions of the *Euronews* test set audio are shown. The X-axis represent the number of OOV PNs selected from the diachronic corpus. Y-axis represent recall (top) and MAP (bottom) of the target OOV PNs. For comparison, Figure 2 shows the retrieval results obtained with 300 topics.

6.1. What do Recall and MAP indicate

It is necessary to understand the importance and differences for the recall and MAP curves. After retrieval of the relevant OOV PNs, the top-N relevant OOV PNs are to be used for recovery/recognition of the target OOV PNs. To recover the target OOV PNs one can use phone matching (Pan et al., 2005), or additional speech recognition pass (Oger et al., 2008); or spotting PNs in speech (Parada et al., 2010). In each of these approaches, the retrieval ranks/scores may or may not be used. This is where the recall and MAP curves make a difference. The recall value at an *operating point* (i.e. N in the top-N choice) will be same whether the ranks of the retrieved OOV PNs are closer to 1 or to N. Whereas the MAP value is a function of the retrieval ranks. It can be seen that the MAP for LX is best but it does not have the best recall rates. This is because with smaller number of OOV PNs to choose from it makes smaller retrieval errors but at the same time gives smaller coverage of the target OOV PNs. We can also observe that the recall values for the reference and LVCSR transcriptions appear close but the MAP values capture the differences in ranks. Similarly, in Figure 1 the recall values for LX and Fig do not improve beyond 300 topics but the MAP values increase due to the improvement in ranks of the target OOV PNs within the top 100 retrieved OOV PNs.

6.2. Effects of different Diachronic Corpora

By analysing the recall and MAP curves of Figure 2 we can draw the following conclusions.

- Expanding time period of diachronic corpus (as in LX-18m) gives better target OOV PN coverage, but not the best recall rates. Additionally it gives a low MAP. Such corpora can be possibly exploited by training more number of topics or by better retrieval methods (Sheikh et al., 2015c; Sheikh et al., 2016).

- Adding documents for less frequent OOV PNs is not effective (performance of LX+rFIG is similar to that of LX+FIG). We found that adding documents for less frequent OOV PNs in LX+rFIG leads to inclusion of more than 60% of FIG documents. The additional

data not only increases data for learning topic representation of less frequent OOV PNs but also comes with additional less frequent OOV PNs and more instances for frequent OOV PNs. Further analysis of the ranks of the less frequent OOV PNs obtained with LX, LX+FIG and LX+rFIG showed that the ranks with LX+rFIG are better with respect to LX but similar to that with LX+FIG.

- Using documents of same time period and from multiple sources (LX+FIG) gives a good balance of recall, MAP and target OOV PN coverage.

7. Conclusion

Diachronic text corpora are essential for recovery of OOV words and proper names missed by LVCSR systems. We performed a study to analyse the selection of text documents in the diachronic corpus used for retrieval of OOV proper names. French broadcast news videos from a particular time period were used as the test set and text articles from different French news websites formed the diachronic text corpora. With OOV proper name retrieval based on topic context, we can conclude that (a) text from a longer time span can give increased coverage of OOV proper names (b) but corpus with text from different sources leads to better retrieval performance than relying on text from single source, even if it corresponds to longer time span (c) and less frequent OOV proper names need improvement in retrieval methods (Sheikh et al., 2016) and not just additional training data.

Acknowledgment

This work is funded by the ContNomina project supported by the French National Research Agency (ANR) under the contract ANR-12-BS02-0009.

8. Bibliographical References

Allauzen, A. and Gauvain, J.-L. (2005). Open vocabulary ASR for audiovisual document indexation. In *IEEE ICASSP*, pages 1013–1016.

Béchet, F., Nasr, A., and Genet, F. (2000). Tagging unknown proper names using decision trees. In *38th Annual Meeting on Association for Computational Linguistics*, pages 77–84, PA, USA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Illina, I., Fohr, D., Mella, O., and Cerisara, C. (2004). The Automatic News Transcription System: ANTS some Real Time experiments. In *INTERSPEECH*, pages 377–380.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Oger, S., Linarès, G., Béchet, F., and Nocera, P. (2008). On-demand new word learning using world wide web. In *IEEE ICASSP*, pages 4305–4308.

Palmer, D. and Ostendorf, M. (2005). Improving out-of-vocabulary name resolution. *Computer Speech & Language*, 19:107 – 128.

Pan, Y.-C., Liu, Y.-Y., and Lee, L.-S. (2005). Named entity recognition from spoken documents using global evidences and external knowledge sources with applications on mandarin chinese. In *IEEE Workshop ASRU*, pages 296–301.

Parada, C., Sethy, A., Dredze, M., and Jelinek, F. (2010). A spoken term detection framework for recovering out-of-vocabulary words using the web. In *INTERSPEECH*, pages 1269–1272.

Parada, C., Dredze, M., and Jelinek, F. (2011). OOV sensitive named-entity recognition in speech. In *INTERSPEECH*, pages 2085–2088.

Qin, L. (2013). *Learning Out-of-Vocabulary Words in Automatic Speech Recognition*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sheikh, I., Illina, I., Fohr, D., and Linares, G. (2015a). OOV proper name retrieval using topic and lexical context models. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5291–5295, April.

Sheikh, I. A., Illina, I., and Fohr, D. (2015b). Study of entity-topic models for OOV proper name retrieval. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1344–1348.

Sheikh, I. A., Illina, I., Fohr, D., and Linarès, G. (2015c). Learning to retrieve out-of-vocabulary words in speech recognition. *CoRR*, abs/1511.05389.

Sheikh, I., Illina, I., Fohr, D., and Linarès, G. (2016). Document level semantic context for retrieving OOV proper names. In *(To Appear) 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.